REVIEW

The Genome Sequence of *Drosophila melanogaster*

Mark D. Adams,[1]* Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[2] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[6] Catherine R. Nelson,[4] George L. Gabor Miklos,[7] Josep F. Abril,[8] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[9] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[2] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[3] Peter Brokstein,[4] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[3] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[3] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[5] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[9] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[3] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[3] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[1] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[6] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[6] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[3] David L. Nelson,[3] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Sidén-Kiamos,[11] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Aihui H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[3] Jean Weissenbach,[14] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[19] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[3] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[34] J. Craig Venter[1]

The fly *Drosophila melanogaster* is one of the most intensively studied organisms in biology and serves as a model system for the investigation of many developmental and cellular processes common to higher eukaryotes, including humans. We have determined the nucleotide sequence of nearly all of the ~120-megabase euchromatic portion of the *Drosophila* genome using a whole-genome shotgun sequencing strategy supported by extensive clone-based sequence and a high-quality bacterial artificial chromosome physical map. Efforts are under way to close the remaining gaps; however, the sequence is of sufficient accuracy and contiguity to be declared substantially complete and to support an initial analysis of genome structure and preliminary gene annotation and interpretation. The genome encodes ~13,600 genes, somewhat fewer than the smaller *Caenorhabditis elegans* genome, but with comparable functional diversity.

The annotated genome sequence of *Drosophila melanogaster*, together with its associated biology, will provide the foundation for a new era of sophisticated functional studies (*1*–*3*). Because of its historical importance, large research community, and powerful research tools, as well as its modest genome size, *Drosophila* was chosen as a test system to explore the applicability of whole-genome shotgun (WGS) sequencing for large and complex eukaryotic genomes (*4*). The groundwork for this project was laid over many years by the fly research community, which has molecularly characterized ~2500 genes; this work in turn has been supported by nearly a century of genetics (*5*). Since *Drosophila* was chosen in 1990 as one of the model organisms to be studied under the auspices of the federally funded Human Genome Project, genome projects in the United States, Europe, and Canada have produced a battery of genome-wide resources (Table 1). The Berkeley and European *Drosophila* Genome Projects (BDGP and EDGP) initiated genomic sequencing (Tables 1 to 3) and finished 29 Mb. The bacterial artificial chromosome (BAC) map and other genomic resources available for *Drosophila* serve both as an independent confirmation of the assembly of data from the shotgun strategy and as a set of resources for further biological analysis of the genome.

The *Drosophila* genome is ~180 Mb in size, a third of which is centric heterochromatin (Fig. 1). The 120 Mb of euchromatin is on two large autosomes and the X chromosome; the small fourth chromosome contains only ~1 Mb of euchromatin. The heterochromatin consists mainly of short, simple sequence elements repeated for many megabases, occasionally interrupted by inserted transposable elements, and tandem arrays of ribosomal RNA genes. It is known that there are small islands of unique sequence embedded within heterochromatin—for example, the mitogen-activated protein kinase gene *rolled* on chromosome 2, which is flanked on each side by at least 3 Mb of heterochromatin. Unlike the *C. elegans* genome, which can be completely cloned in yeast artificial chromosomes (YACs), the simple sequence repeats are not stable in YACs (*6*) or other large-insert cloning sys-

tems. This has led to a functional definition of the euchromatic genome as that portion of the genome that can be cloned stably in BACs. The euchromatic portion of the genome is the subject of both the federally funded *Drosophila* sequencing project and the work presented here. We began WGS sequencing of *Drosophila* less than 1 year ago, with two major goals: (i) to test the strategy on a large and complex eukaryotic genome as a prelude to sequencing the human genome, and (ii) to provide a complete, high-quality genomic sequence to the *Drosophila* research community so as to advance research in this important model organism.

WGS sequencing is an effective and efficient way to sequence the genomes of prokaryotes, which are generally between 0.5 and 6 Mb in size (*7*). In this strategy, all the DNA of an organism is sheared into segments a few thousand base pairs (bp) in length and cloned directly into a plasmid vector suitable for DNA sequencing. Sufficient DNA sequencing is performed so that each base pair is covered numerous times, in fragments of 500 bp. After sequencing, the fragments are assembled in overlapping segments to reconstruct the complete genome sequence.

In addition to their much larger size, eukaryotic genomes often contain substantial amounts of repetitive sequence that have the potential to interfere with correct sequence assembly. Weber and Myers (*8*) presented a theoretical analysis of WGS sequencing in which they examined the impact of repetitive sequences, discussed experimental strategies to mitigate their effect on sequence assembly, and suggested that the WGS method could be applied effectively to large eukaryotic genomes. A key component of the strategy is obtaining sequence data from each end of the cloned DNA inserts; the juxtaposition of these end-sequences ("mate pairs") is a critical element in producing a correct assembly.

Genomic Structure

WGS libraries were prepared with three different insert sizes of cloned DNA: 2 kb, 10 kb, and 130 kb. The 10-kb clones are large enough to span the most common repetitive sequence elements in *Drosophila*, the retrotransposons. End-sequence from the BACs provided long-range linking information that was used to confirm the overall structure of the assembly (*9*). More than 3 million sequence reads were obtained from whole-genome libraries (Fig. 2 and Table 2). Only 2% of the sequence reads contained heterochromatic simple sequence repeats, indicating that the heterochromatic DNA is not stably cloned in the small-insert vectors used for the WGS libraries. A BAC-based physical map spanning >95% of the euchromatic portion of the genome was constructed by screening a BAC library with sequence-tagged site (STS) markers (*10*). More than 29 Mb of high-quality finished sequence has been completed from BAC, P1, and cosmid clones, and draft sequence data (1.5× average coverage) were obtained from an additional 825 BAC and P1 clones spanning in total >90% of the genome (Table 3). The clone-based draft sequence served two purposes: It improved the likelihood of accurate assembly, and it allowed the identification of templates and primers for filling gaps that remain after assembly. An initial assembly was performed using the WGS data and BAC end-sequence [WGS-only assembly (*4*)]; subsequent assemblies included the clone-based draft sequence data (joint assembly). Figure 3 and Table 3 illustrate the status of the euchromatic sequence resulting from each of these assemblies and the current status following the directed gap closure completed to date. The sequence assembly process is described in detail in an accompanying paper (*11*).

Assembly resulted in a set of "scaffolds." Each scaffold is a set of contiguous sequences (contigs), ordered and oriented with respect to one another by mate-pairs such that the gaps between adjacent contigs are of known size and are spanned by clones with end-sequences flanking the gap. Gaps within scaffolds are called sequence gaps; gaps between scaffolds are called "physical gaps" because there are no clones identified spanning the gap. Two methods were used to map the scaffolds to chromosomes: (i) cross-referencing between STS markers present in the assembled sequence and the BAC-based STS content map, and (ii) cross-referencing between assembled sequence and shotgun sequence data obtained from individual tiling-path clones selected from the BAC physical map. The mapped scaffolds from the joint assembly, totaling 116.2 Mb after initial

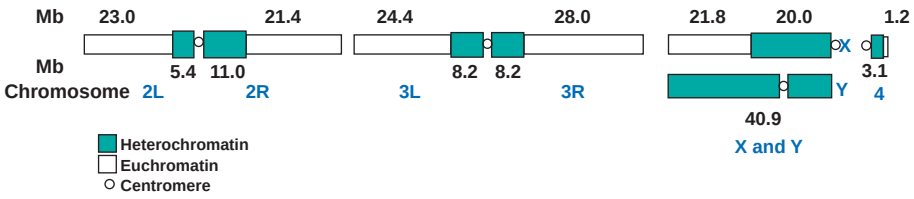

Fig. 1. Mitotic chromosomes of *D. melanogaster*, showing euchromatic regions, heterochromatic regions, and centromeres. Arms of the autosomes are designated 2L, 2R, 3L, 3R, and 4. The euchromatic length in megabases is derived from the sequence analysis. The heterochromatic lengths are estimated from direct measurements of mitotic chromosome lengths (*67*). The heterochromatic block of the X chromosome is polymorphic among stocks and varies from one-third to one-half of the length of the mitotic chromosome. The Y chromosome is nearly entirely heterochromatic.

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]Berkeley *Drosophila* Genome Project (BDGP), Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [4]BDGP, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [5]European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [6]Neomorphic Inc., 2612 Eighth Street, Berkeley, CA 94710, USA. [7]GenetixXpress Pty. Ltd., 78 Pacific Road, Palm Beach, Sydney, NSW 2108, Australia. [8]Department of Medical Informatics, IMIM–UPF C/Dr. Aiguader 80, 08003 Barcelona, Spain. [9]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. [10]Department of Genetics, Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, MO 63110, USA. [11]Institute of Molecular Biology and Biotechnology, Forth, Heraklion, Greece. [12]European *Drosophila* Genome Project (EDGP), EMBL, Heidelberg, Germany. [13]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94710, USA. [14]Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France. [15]Section of Molecular and Cellular Biology, University of California, Davis, CA 95618, USA. [16]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [17]EDGP, Rennes University Medical School, UPR 41 CNRS Recombinaisons Genetiques, Faculte de Medicine, 2 av. du Pr. Leon Bernard, 35043 Rennes Cedex, France. [18]Department of Genetics, Stanford University, Palo Alto, CA 94305, USA. [19]Department of Statistics, University of California, Berkeley, CA 94720, USA. [20]EDGP, Centro de Biología Molecular Severo Ochoa, CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain. [21]MBVL, Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA. [22]Department of Biochemistry and Center for Insect Science, University of Arizona, Tucson, AZ 85721, USA. [23]EDGP, Montpellier University Medical School, Institut de Genetique Humaine, CNRS (CRBM), 114 rue de la Cardonille, 34396 Montpellier Cedex 5, France. [24]Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health (NIH), Bethesda, MD 20892, USA. [25]Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, Canada. [26]The Institute for Genomic Research, Rockville, MD 20850, USA. [27]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA. [28]EDGP, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [29]Department of Biochemistry, University of Tennessee, Memphis, TN 38163, USA. [30]EDGP, Department of Anatomy and Physiology, University of Dundee, Dundee DD1 4HN, UK, and Department of Biological Sciences, Open University, Milton Keynes MK7 6AA, UK. [31]HHMI/Embryology, Carnegie Institution of Washington, Baltimore, MD 21210, USA. [32]Motorola BioChip Systems, Tempe, AZ 85284, USA. [33]Cell Biology and Metabolism Branch, National Institute of Child Health and Human Development, NIH, Bethesda, MD 20892, USA. [34]Howard Hughes Medical Institute, BDGP, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed.

gap closure, were deposited in GenBank (accession numbers AE002566–AE003403) and form the basis for the analysis described in this article.

The WGS-only assembly resulted in 50 scaffolds spanning 114.8 Mb that could be placed unambiguously onto chromosomes solely on the basis of their STS content (labeled "D" in Fig. 3). The joint assembly included clone-based sequence, but no specific advantage was taken of the location information of each clone-based read by the whole-genome assembly algorithm. Nonetheless, the clone-based sequence from BACs in the physical map allowed placement of an additional 84 small scaffolds (1.4 Mb) on chromosome arms in the joint assembly (labeled "C" in Fig. 3). As shown in Fig. 3, a few large scaffolds in each assembly span a large portion of each chromosome arm, with a number of additional smaller scaffolds located at the centromeric end, except on the right arm of chromosome 3. Nearly all of the scaffolds added to chromosomes in the joint assembly, relative to the WGS-only assembly, are adjacent to the centric heterochromatin, which demonstrates the utility of the physical map in these regions. The density of transposable elements (labeled "A" in Fig. 3) increases markedly in the transition zone between euchromatin and heterochromatin, as discussed below. An additional 704 scaffolds in the joint assembly, equivalent to 3.8 Mb, could not be placed with accuracy on the genome. Most of these do not match clone-based sequence from the physical map, and therefore they most likely represent small islands of unique sequence embedded within regions of heterochromatin. Because of the instability of the surrounding genomic regions, these sequences would not have been obtained through a sequencing approach that was dependent on cloning in large-insert vectors.

Among the 134 mapped scaffolds, there were 1636 contigs after assembly (hence 1630 gaps, considering that there are six linear chromosome arm segments to be assembled). On the major autosomes, there are five physical gaps in the BAC map, three of which are near a centromere or telomere (*10*). Because the WGS approach did not span these gaps, they likely contain unclonable regions. Most gaps on the autosomes—including gaps between scaffolds—were therefore cloned in either WGS clones or BAC subclones used for clone-based draft sequencing and are considered sequence gaps. Directed gap closure was done through use of several resources, including whole BAC clones, plasmid subclones, and M13 subclones from the Lawrence Berkeley National Laboratory (LBNL) and Baylor College of Medicine centers' draft sequence of BAC and P1 clones; 10-kb subclones from the whole-genome libraries; and polymerase chain reaction (PCR) from genomic DNA (*12*). The average size of the gaps filled to date is 771 bp (their predicted size was 757 bp); the predicted size of the remaining gaps is 2120 bp. Table 3 provides details of the status of each chromosome arm as of 3 March 2000.

The accuracy of the assembly was measured in several ways, as described (*11*). In summary, the scaffold sequences agree very well with the BAC-based STS content map and with high-quality finished sequence. In the 7 Mb of the genome where very high-quality sequence was

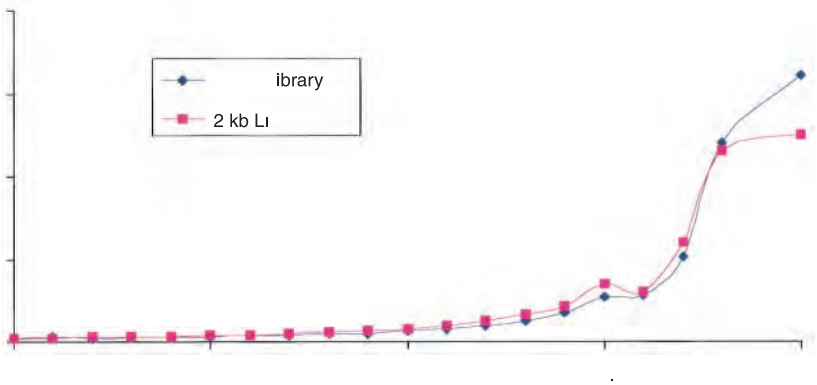

Fig. 2. Accuracy of sequence reads from ABI Prism 3700 DNA analyzer. A database of BAC and P1 clone sequences from BDGP finished to high accuracy ($P_{sum} > 100{,}000$, indicating less than one error predicted per 100,000 bases) was constructed. Trimmed WGS sequence reads matching these BAC and P1 clones were identified by BLAST. The first high-scoring pair (HSP) with a full-length match was used. Identity is the percentage of matched nucleotides in the alignment; 49,756 sequence reads from 2-kb libraries and 23,455 reads from 10-kb libraries matched these BAC and P1 sequences. The average trimmed read length of sequences from 2-kb and 10-kb clones was 570 bp and 567 bp, respectively.

Table 1. Genomic resources for *Drosophila*.

Type	Description	Resolution	Contribution	Source and reference
BAC-based STS content map	STS content map constructed by screening 23× genome coverage of BAC clones; a tiling path of BACs spanning each chromosome arm was selected	50 kb	Location of whole-genome scaffolds to chromosomes; confirmation of accuracy of assembly	BDGP [chromosomes 2 and 3 (*10*)], EDGP [X chromosome (*69*), www.dundee.ac.uk/anatphys/robert/Xdivs/MapIntro.htm], University of Alberta [chromosome 4 (*70*)]
Polytene map	Tiling-path BACs hybridized to polytene chromosomes	30 kb	Location of STSs and BACs to chromosomes; validation of BAC map	See (*10*)
BAC end-sequence	500 bp of sequence from each end of a BAC clone	Two reads per 130 kb	Long-range association of sequence contigs	Genoscope (www.genoscope.fr)
Finished clone-based sequence	BAC, P1, and cosmid clones completely sequenced to high accuracy	29 Mb of total sequence	Assessment of accuracy of Celera sequence and assembly	LBNL (26 Mb), EDGP [3 Mb (*69*)]
Draft sequence from mapped BACs	≥1.5× shotgun sequence coverage of 825 clones from the tiling path of BAC and P1 clones	384 reads distributed across 160 kb	Location of sequence contigs to a small genomic region; templates for gap closure	LBNL, Baylor College of Medicine

available for comparison, the accuracy of the assembled sequence was 99.99% in nonrepetitive regions. In the 2.5% of the region comprising the most highly repetitive sequences, the accuracy was 99.5%.

Heterochromatin-euchromatin transition zone. The genomes of eukaryotes generally contain heterochromatic regions surrounding the centromeres that are intractable to all current sequencing methods. In *Drosophila*, 60 Mb of the 180-Mb genome consists of centric heterochromatin, which is composed primarily of simple sequence satellites, transposons, and two large blocks of ribosomal RNA genes (*13*). We examined the sequence organization at boundaries between euchromatin and centric heterochromatin in two regions, one in division 20 on the X chromosome and the other in division 40 on the left arm of chromosome 2. On the X chromosome, gene density in division 20 drops abruptly—to two genes in 400 kb around *folded gastrulation*—and then rises to 11 genes in 130 kb. Next, at least 10 Mb of largely satellite DNA sequences and the ribosomal RNA gene cluster are located just distal to the centromere itself. On the left arm of chromosome 2, a similar situation exists: There is a normal gene density in division 39, followed by only two genes in 350 kb near *teashirt* in division 40, then by a 200-kb region containing 10 genes. These transition zones between euchromatin and heterochromatin contain many previously unknown genes, including counterparts to human cyclin K and mouse Krox-4. None of the 11 genes proximal to *teashirt* and only one of the 10 genes proximal to *folded gastrulation* was known previously.

What is the nature of the sequence in the gene-poor regions? The most common sequences by far were transposons, consistent with previous small-scale analyses (*14*). These include several new elements similar to transposons in other species, as well as the 50 transposon classes previously characterized in *Drosophila*. Some short runs of satellite sequences are present, but it has not been determined whether they might have been truncated during cloning. In addition, at least 110 other simple repeat classes were identified, some of which are distributed widely outside of heterochromatin.

Criteria for describing the completion status of a eukaryotic genome. Because of the unclonable repetitive DNA surrounding the centromeres, it is highly unlikely that the genomic sequence of chromosomes from eukaryotes such as *Drosophila* or human will ever be "complete." It is therefore necessary to provide an assessment of the contiguity and accuracy of the sequence. Table 4 lists several objective parameters by which the status can be judged and by which improvements in future releases can be measured. We have termed the version of the sequence associated with this publication "Release 1" and intend to make regular future releases as gaps are filled and overall sequence accuracy is increased.

One measure of the completeness of the assembled sequence is the extent to which previously described genes can be found. An analysis of the 2783 *Drosophila* genes with some sequence information that have been compiled by FlyBase (*15*) resulted in identification of 2778 on the scaffold sequence. All of the remainder are found in unscaffolded sequence. The remaining six were all cloned by degenerate PCR, and it is possible that some or all of these genes are incorrectly ascribed to *Drosophila* (*16*). Of the base pairs represented in the 2778 genes, 97.5% are present in the assembled sequence.

Annotation

The initial annotation of the assembled genome concentrated on two tasks: prediction of transcript and protein sequence, and prediction of function for each predicted protein. Computational approaches can aid each task, but biologists with expertise in particular fields are required for the results to have the most consistency, reliability, and utility. Because the breadth of expertise necessary to annotate a complete genome does not exist in any single individual or organization, we hosted an "Annotation Jamboree" involving more than 40 scientists from around the world, primarily from the *Drosophila* research community. Each was responsible for organizing and interpreting the gene set for a given protein family or biological process. Over a 2-week period, jambo

Table 2. Source of data for assembly: Whole-genome shotgun sequencing. See (*65*) for more information about library construction and sequencing.

Vector	Insert size (kbp)	Paired sequences	Total sequences	Clone coverage	Sequence coverage
High-copy plasmid	2	732,380	1,903,468	11.2×	7.3×
Low-copy plasmid	10	548,974	1,278,386	42.2×	5.4×
BAC	130	9,869	19,738	11.4×	0.07×
Total		1,290,823	3,201,592	64.8×	12.8×

Table 3. BAC and P1 clone-based sequencing. EDGP, European *Drosophila* Genome Project; BCM, Baylor College of Medicine; LBNL, Lawrence Berkeley National Laboratory (BCM and LBNL are the genomic sequencing centers of the BDGP).

Clone-based genomic sequencing								Gap closure: current status		
Chromosomal region	Group	Size	Finished sequence (Mb)	Draft sequence in joint assembly [BACs, (P1s)]†		Total sequenced BACs (P1s) in joint assembly	Additional sequenced BACs in tiling path	Percentage of DNA sequence in contigs greater than		
				Clones	Average coverage			30 kb	100 kb	1 Mb
X (1–3)	EDGP	3	2.5	0		0	0	79.4	32.7	0
X (4–11)	BCM	8.8	0.1*	0		1	72			
X (12–20)	LBNL	10	0	71	2.3×	71	10			
2L	LBNL	23	14.0	103 (8)	1.6× (5.3×)	119 (202)	2	97.8	91.4	16.9
2R	LBNL	21.4	8.8	159 (32)	1.3× (4.7×)	157 (186)	0	96.4	90.6	32.8
3L	BCM	24.4	0.1	166	1.3×	170	50	95.1	77.7	0
3L	LBNL		2.1	22 (7)	1.7× (2.5×)	20 (32)	0			
3R	LBNL	28	2.1	259 (9)	1.2× (2×)	264 (27)	0	98.5	92.6	3.6
4	LBNL	1.2	0	16	1.4×	15	1	85.6	43.5	0
Total		120	29.7	796 (56)		817 (447)	135	93.7	77.5	9.9

*Sequenced at LBNL. †A tiling path of clones spanning 97% of the euchromatic portion of the genome was selected from the genome physical maps (*10*) for clone-based sequencing. The data include sequence that has been generated since the beginning of the publicly funded (BDGP and EDGP) genome sequencing projects. Tiling path clone identities were verified by screening the shotgun sequence for expected STS and BAC end-sequences, sequenced genes with known map locations from genes (and regions flanking P insertions), and sequences of neighboring tiling path clones. The average size of BAC clones in the tiling path is 163 kb. Sequencing methods are described in (*66*).

Fig. 3. Assembly status of the *Drosophila* genome. Each chromosome arm is depicted with information on content and assembly status: (**A**) transposable elements, (**B**) gene density, (**C**) scaffolds from the joint assembly, (**D**) scaffolds from the WGS-only assembly, (**E**) polytene chromosome divisions, and (**F**) clone-based tiling path. Gene density is plotted in 50-kb windows; the scale is from 0 to 30 genes per 50 kb. Gaps between scaffolds are represented by vertical bars in (C) and (D). Clones colored red in the tiling path have been completely sequenced; clones colored blue have been draft-sequenced. Gaps shown in the tiling path do not necessarily mean that a clone does not exist at that position, only that it has not been sequenced. Each chromosome arm is oriented left to right, such that the centromere is located at the right side of X, 2L, and 3L and the left side of 2R and 3R.

ree participants worked to define genes, to classify them according to predicted function, and to begin synthesizing information from a genome-wide perspective.

For definition of gene structure, we relied on the use of different gene-finding approaches: the gene-finding programs Genscan (*17*) and a version of Genie that uses expressed sequence tag (EST) data (*18*), plus the results of complementary DNA (cDNA) and protein database searches, followed by review by human annotators (*19*). Genscan predicted 17,464 genes, and Genie predicted 13,189. We believe that the lower estimate is more accurate, because in a test that used the extensively studied and annotated 2.9-Mb *Adh* region (*3*), the Genie predictions were closer to the number of experimentally determined genes; Genscan predicted far too many (*20*). This is likely because Genie was optimized for *Drosophila*, whereas Genscan parameters suitable for *Drosophila* gene-finding are not available.

Results of the computational analyses were presented to annotators by means of a custom visualization tool that allowed annotators to define transcripts on the basis of EST (*21*) and protein sequence similarity information, Genie predictions, and Genscan predictions, in decreasing order of confidence. The present annotation of the *Drosophila* genome predicts 13,601 genes, encoding 14,113 transcripts through alternative splicing in some genes. The number of alternative splice forms that can be annotated is limited by the available cDNA data and is a substantial underestimate of the total number of alternatively spliced genes. More than 10,000 genes with database matches were reviewed manually. The remaining 3000 genes were predicted by Genie but have no database matches that can be used to refine intron-exon boundaries. Genes predicted by Genscan that did not overlap Genie predictions or database matches were not included in the set of predicted proteins. Table 5 summarizes the evidence for these genes: 38% of the Genie predictions are supported by evidence from both EST and protein matches, 27% by ESTs alone, and 12% by protein matches alone. Altogether there are EST matches for 65% of the genes, but nearly half of the total ESTs match only 5% of the genes; 23% of the predicted proteins do not match sequences from other organisms or *Drosophila* ESTs. This set of annotations is considered provisional and will improve as additional full-length cDNA sequence and functional information becomes available for each gene. Figure 4 provides a graphical overview of the gene content of the fly.

Genes were classified according to a functional classification scheme called Gene Ontology (GO). The GO project (*22*) is a collaboration among FlyBase, the *Saccharomyces* Genome Database (*23*), and Mouse Genome Informatics (*24*). It consists of a set of controlled vocabularies providing a consistent description of gene products in terms of their molecular function, biological role, and cellular location. At the time of our annotation, proteins encoded by 1539 *Drosophila* genes had already been annotated by FlyBase using 1200 different GO classifications. In addition, a set of 718 proteins from *S. cerevisiae* and 1724 proteins from mouse had been annotated and placed into GO categories. Predicted *Drosophila* genes and gene products were used as queries against a database made up of the sequences of these three sets of proteins (by BLASTX or BLASTP) (*25*) and grouped on the basis of the GO classification of the proteins matched. About 7400 transcripts have been assigned to 39 major functional categories, and about 4500 have been assigned to 47 major process categories (Table 6).

The largest predicted protein is Kakapo, a cytoskeletal linker protein required for adhesion between and within cell layers, with 5201 amino acids; the smallest is the 21–amino acid ribosomal protein L38. There are 56,673 predicted exons, an average of four per gene, occupying 24.1 Mb of the 120-Mb euchromatic sequence total. The size of the average predicted transcript is 3058 bp. There was a systematic underprediction of 5′ and 3′ untranslated sequence as a result of less than complete EST coverage and the inability of gene-prediction programs to predict the noncoding regions of transcripts, so the number of exons and introns and the average transcription unit size are certain to be underestimates. There are at least 41,000 introns, occupying 20 Mb of sequence. Intron sizes in *Drosophila* are heterogeneous, ranging from 40 bp to more than 70 kb, with a clear peak between 59 and 63 bp (*26*). The average number of exons is four, although this is an underestimate because of a systematic underprediction of 5′ and 3′ untranslated exons. We identified 292 transfer RNA genes and 26 genes for spliceosomal small nuclear RNAs (snRNAs). We did not attempt to predict other noncoding RNAs.

The total number of protein-coding genes, 13,601, is less than that predicted for the worm *C. elegans* (*27*) (18,425; WormPep 18, 11 October 1999) and far less than the 27,000 estimated for the plant *Arabidopsis thaliana* (*28*). The average gene density in *Drosophila* is one gene per 9 kb. There is substantial variation in gene density, ranging from 0 to nearly 30 genes per 50 kb, but the gene-rich regions are not clustered as they are in *C. elegans*. Regions of high gene density correlate with G+C-rich sequences. In the 1 Mb adjacent to the centric heterochromatin, both G+C content and gene density decrease, although there is not a marked decrease in EST coverage as has been seen in *A. thaliana* (*28*).

Table 4. Measures of completion. Analyses supporting many of these values are found in (*11*).

Number of scaffolds mapped to chromosome arms	134
Number of scaffolds not mapped to chromosomes	704
Number of base pairs in scaffolds mapped to chromosome arms	116.2 Mb
Number of base pairs in scaffolds not mapped to chromosome arms	3.8 Mb
Largest unmapped scaffold	64 kb
Percentage of total base pairs in mapped scaffolds >100 kb	98.2%
Percentage of total base pairs in mapped scaffolds >1 Mb	95.5%
Percentage of total base pairs in mapped scaffolds >10 Mb	68.0%
Number of gaps remaining among mapped scaffolds	1299
Base pair accuracy against LBNL BACs (nonrepetitive sequence)	99.99%
Known genes accounted for in scaffold set	99.7%

Genomic Content

The genomic sequence has shed light on some of the processes common to all cells, such as replication, chromosome segregation, and iron metabolism. There are also new findings about important classes of chromosomal proteins that allow insights into gene regulation and the cell cycle. Overall, the correspondence of *Drosophila* proteins involved in gene expression and metabolism to their human counterparts reaffirms that the fly represents a suitable experimental platform for the examination of human disease networks involved in replication, repair, translation, and the metabolism of drugs and toxins. In an accompanying manuscript (*29*), the protein complement of *Drosophila* is compared to those of the two eukaryotes with complete genome sequences, *C. elegans* and *S. cerevisiae*, and other developmental and cell biological processes are discussed.

Replication. Genes encoding the basic DNA replication machinery are conserved among eukaryotes (*30*); in particular, all of the proteins known to be involved in start site recognition are encoded by single-copy genes in the fly. These include members of the six-subunit heteromeric origin recognition complex (ORC) (*31*), the MCM helicase complex (*32*), and the regulatory factors CDC6 and CDC45, which are thought to determine processing of pre-initiation complexes. The fly ORC3 and ORC6 proteins, for example, share close sequence similarity with vertebrate proteins, but not only are they highly divergent relative to yeast ORCs, they have no obvious counterparts in the worm. It is striking that the ORC genes exist as single copies, given the orthologous functions for some of the subunits in other processes (*33*). It had been considered possible that a large family of ORCs, each with a different binding specificity, might account for

different origin usage in development. Clearly, given the single-copy ORC genes, other as-yet-undiscovered cis-acting elements and trans-acting factors participate in developmentally regulated processes such as switches in origin usage, gene amplification, and specialized replication of euchromatin in certain endocycles. In contrast, the fly has two distinct homologs of the proliferating cell nuclear antigen (PCNA), the processivity factor for the DNA polymerases (δ and ε) involved in chain elongation. Human PCNA is blocked from interaction with the replication enzymes by the checkpoint regulator p21 in response to DNA damage (*34*); perhaps one of the fly PCNA proteins is immune to such regulation and is thus left active for repair or replication.

Chromosomal proteins. Analysis of protein families involved in chromosome inheritance reveals both expected findings and some surprises. As expected, the fly has all four members of the conserved SMC family involved in sister chromatid cohesion, condensation, DNA repair, and dosage compensation (*35*). The fly also contains at least one ortholog of each of the MAD/Bub metaphase-anaphase checkpoint proteins that are conserved from yeast to mammals. However, *Drosophila* does not appear to have orthologs to most of the proteins identified previously in mammals or yeast that are associated with centromeric DNA, such as the CENP-C/MIF-2 family and the yeast CBF3 complex (*36*). One exception is the presence of a histone H3-like protein that shares sequence similarity with mammalian CENP-A, a centromere-specific H3-like protein. There are at least nine histone acetyltransferases (HATs) and five histone deacetylases (HDACs), which are involved in regulating chromatin structure (*37*); only three of each have been reported previously. There are also 17 members of the SNF2 adenosine triphosphatase (ATPase) family, which represent 9 of the 10 known subfamilies. Many of these ATPases are involved in chromatin remodeling (*38*). The fly also contains at least 14 proteins with chromodomains (*39*), six of which are new, including two HP1-related proteins. Although many of these chromodomain-containing proteins have orthologs in vertebrates, only one (CHD1) appears in yeast, flies, and vertebrates. There are also at least 13 bromodomain-containing proteins, seven of which are new; the bromodomain may interact with the acetylated NH_2-terminus of histones and is involved in chromatin remodeling and gene silencing (*40*). Only three of these appear to have counterparts in yeast. Furthermore, *Drosophila* telomeres lack the simple repeats that are characteristic of most eukaryotic telomeres (*41*), and the known telomerase components of vertebrates, for example, are absent from flies. The fly does, however, contain five proteins that are close relatives of the yeast and human SIR2 telomere silencing proteins.

DNA repair. The importance of DNA repair in maintaining genomic integrity is reflected in the conservation of most proteins implicated in the major defined pathways of eukaryotic DNA repair. However, there are some notable absences. For example, no convincing homologs can be found for the genes encoding the RAD7, RAD16, RAD26 (CSB/ERCC6), and RAD28 (CSA) proteins, which are implicated in strand-specific modes of repair in yeast and/or mammalian systems. In base excision repair processes, 3-methyladenine glycosylase and uracil-DNA-glycosylase are absent, although the latter function is likely fulfilled by the G/T mismatch-specific thymine DNA glycosylase (*42*). In the damage bypass pathway, sequences encoding homologs of DNA polymerase ζ (yeast Rev3p/*Drosophila* mus205) and Rev1p are present, although a REV7 homolog is not found. As in humans and worms, two members of the RAD30 (polymerase η) gene family are present. In the mismatch repair system, only two proteins related to *Escherichia coli* mutS are predicted, rather than the usual family of five or more members. The previously reported Msh2p homolog (*43*) is present, as is a sequence most closely resembling Msh6p. Budding yeast and humans possess additional members of the mutS gene family that are proposed to function in partially redundant pathways of mismatch repair (MSH3) and in meiotic recombination (MSH4 and MSH5), suggesting either that the *Drosophila* mutS homologs have reduced specificity or that alternative proteins are fulfilling these roles in the fly. In the recombinational repair pathway, two additional members of the recA/RAD51 gene family are identified, bringing the total to four. However, no member of the RAD52/RAD59 family is present. One additional member of the recQ/SGS1 helicase family was identified, in addition to the two already noted (*44*); the new protein is most similar to human RecQ4. Finally, with respect to nonhomologous end joining, *Drosophila* joins the list of invertebrate species that lack an apparent DNA-PK catalytic subunit, although both Ku subunits and DNA ligase 4 are present. We conclude that most major components of the repair network in flies have been uncovered. If more are present, either they have diverged so far that they are unrecognizable by BLAST searches, or the systems have become degenerate (that is, other network components are fulfilling the same roles).

Transcription. Gene regulation has traditionally been singled out as one of the primary bases for the generation of evolutionary diversity. How has the core transcriptional machinery changed in different phyla? *Drosophila* core RNA polymerase II and some general transcription factors (TFIIA-H, TFIIIA, and TFIIIB) are similar in composition to those of both mammals and yeast (*45*). In contrast, core RNA polymerases I and III, TBP (TATA-binding protein)–containing complexes for class I, class II, and snRNA genes (TBP-associated factors TAF_I and TAF_{II}, and $SNAP_C$, respectively), TFIIIC, and SRB/mediator vary greatly in composition in *Drosophila* and mammals relative to yeast (*46*). The RNA polymerase I transcription factors of flies and mammals have clear amino acid conservation; yeast RNA polymerase I factors do not appear to be related to them. For example, the mammalian promoter interacting factors UBF and TIF-1A are present in *Drosophila* but not in yeast, and yeast UAF subunits are absent in *Drosophila* and apparently absent in mammals. Furthermore, of the three TAF_Is in the human selectivity factor 1, the mouse transcriptional initiation factor IB, and the yeast core factor complexes, only the human/mouse TAF_I63/TAF_I68 subunit is conserved in the fly. Similarly, *Drosophila* encodes three of the five mammalian $SNAP_C$ subunits (SNAP43, 50, and 190) for which no homologs exist in the yeast genome.

In addition to the family of previously described TBPs (*47*), the fly contains multiple forms of several ubiquitous TAF_{II}s ($TAF_{II}30\beta$, $TAF_{II}60$, and $TAF_{II}80$) (*46*). This raises the possibility that a variety of TFIID complexes evolved in metazoan organisms to regulate gene expression patterns associated with development and cellular differentiation. The constellation of factors that interact with RNA polymerase II in *Drosophila* may also contribute to this regulation, because *Drosophila* contains only a small subset of yeast SRB/mediator subunits (MED6, MED7, and SRB7) but a vast majority of the molecularly characterized com-

Table 5. Summary of the gene predictions in *Drosophila*. Gene prediction programs were used in combination with searches of protein and EST databases.

Result	Genie + Genscan*	Genie only†	Genscan only‡	No gene prediction§	Total
EST + protein match	6,040	288	239	49	6,616
EST match only	1,357	143	107	34	1,641
Protein match only	2,541	157	220	78	2,996
No match	1,980	307	0	0	2,348
Total	11,918	895	627	161	13,601

*Genie and Genscan matches overlapped but were not necessarily identical. †Genie predictions in regions not predicted by Genscan. ‡Genscan predictions in regions not predicted by Genie; in the absence of database matches, >4000 Genscan predictions were not included in the annotated gene set. §Gene structures defined based on database matches in the absence of gene predictions.

ponents of mammalian coactivator complexes such as ARC/DRIP/TRAP.

Gene regulation. On the basis of similarity to known proteins, *Drosophila* appears to encode about 700 transcription factors, about half of which are zinc-finger proteins. By contrast, the worm has about 500 transcription factors, fewer than one-third of which are zinc-finger proteins (*29*). Two additional classes play key roles in regulation: the homeodomain-containing and nuclear hormone receptor–type transcription factors.

Homeodomain-containing proteins control a wide variety of developmental processes. Twenty-two new homeodomain-containing proteins were uncovered in our analysis, bringing the total to more than 100. Ten of these were members of the paired-box PRX superclass (*48*), some with known vertebrate homologs: short stature homeobox 2 (SHOX), cartilage homeoprotein 1 (CART), and the two retina-specific proteins (VSX-1 and VSX-2) of goldfish. New members were also found in the LIM and TGIF class. The two new LIM members contain a homeobox and two copies of the LIM motif; the two new TGIF members occur as a local tandem duplication on the right arm of chromosome 2. We also found single new members of the NK-2, muscle-specific homeobox, proline-rich homeodomain (PRH), and BarH classes. The new fly gene encoding NK-2 is a cognate of the gene encoding the NKX-5.1 mouse protein. The new fly gene encoding muscle-specific homeobox is most similar to the gene encoding the MSX-1 mouse protein involved in craniofacial morphogenesis. The new fly gene encoding PRH is most similar to a mouse gene expressed in myeloid cells. The remaining homeodomain-containing proteins are orphans: One has similarity to the human H6 protein involved in craniofacial development, and another to HB9, a protein required for normal development of the pancreas.

Nuclear hormone receptors (NRs) are

Table 6. Gene Ontology (GO) classification of *Drosophila* gene products. Each of the 14,113 predicted transcripts was searched by BLAST against a database of proteins from fly, yeast, and mouse that had been assigned manually to a function and/or process category in the GO system. Function categories were reviewed manually, and in many cases a *Drosophila* protein was assigned to a different category upon careful inspection. The number of transcripts assigned to each process category is the result of computational searches only. For functions, the number of transcripts assigned and manually reviewed in each category is shown (with the results of the computational search in parentheses). Certain cases illustrate the value of the manual inspection. For example, motor proteins initially included many coiled-coil domain proteins incorrectly assigned to this category by the computational search. Supplemental data are available at www.celera.com.

Function	Number of transcripts	Process	Number of transcripts
Nucleic acid binding	1387 (1370)	Cell growth and maintenance	3894
DNA binding	919 (652)	Metabolism	2274
DNA repair protein	65 (30)	Carbohydrate metabolism	53
DNA replication factor	38 (18)	Energy pathways	69
Transcription factor	694 (418)	Electron transport	8
RNA binding	259 (205)	Nucleotide and nucleic acid metabolism	1078
Ribosomal protein	128 (116)	DNA metabolism	64
Translation factor	69 (68)	DNA replication	57
Transcription factor binding	21 (116)	DNA repair	110
Cell cycle regulator	52 (104)	DNA packaging	112
Chaperone	159 (158)	Transcription	735
Motor protein	98 (373)	Amino acid and derivative metabolism	69
Actin binding	93 (64)	Protein metabolism	685
Defense/immunity protein	47 (41)	Protein biosynthesis	215
Enzyme	2422 (2021)	Protein folding	52
Peptidase	468 (456)	Protein modification	273
Endopeptidase	378 (387)	Proteolysis and peptidolysis	81
Protein kinase	236 (307)	Protein targeting	51
Protein phosphatase	93 (93)	Lipid metabolism	111
Enzyme activator	9 (19)	Monocarbon compound metabolism	6
Enzyme inhibitor	68 (92)	Coenzymes and prosthetic group metabolism	23
Apoptosis inhibitor	15 (17)	Transport	336
Signal transduction	622 (554)	Ion transport	72
Receptor	337 (336)	Small molecule transport	109
Transmembrane receptor	261 (280)	Mitochondrial transport	43
G protein–linked receptor	163 (160)	Ion homeostasis	8
Olfactory receptor	48 (49)	Intracellular protein traffic	116
Storage protein	12 (27)	Cell death	50
Cell adhesion	216 (271)	Cell motility	9
Structural protein	303 (302)	Stress response	223
Cytoskeletal structural protein	106 (54)	Defense (immune) response	149
Transporter	665 (517)	Organelle organization and biogenesis	417
Ion channel	148 (188)	Mitochondrion organization and biogenesis	5
Neurotransmitter transporter	33 (18)	Cytoskeleton organization and biogenesis	390
Ligand binding or carrier	327 (391)	Cytoplasm organization and biogenesis	7
Electron transfer	124 (117)	Cell cycle	211
Cytochrome P450	88 (84)	Cell communication	530
Ubiquitin	11 (17)	Cell adhesion	228
Tumor suppressor	10 (5)	Signal transduction	279
Function unknown/unclassified	7576 (7654)	Developmental processes	486
Conserved hypothetical	(1474)	Sex determination	7
		Physiological processes	201
		Sensory perception	64
		Behavior	54
		Process unknown/unclassified	8884

sequence-specific, ligand-dependent transcription factors that contribute to physiological homeostasis by functioning as both transcriptional activators and repressors. Examination of the fly genome revealed only four additional NR members, bringing the total to 20. In contrast, the NR family represents the most abundant class of transcriptional regulators in the worm: More than 200 member genes have been described. One of the newly identified fly NRs possesses a new P-box element (Cys-Asp-Glu-Cys-Ser-Cys-Phe-Phe-Arg-Arg), which confers DNA binding specificity, bringing to 76 the number of P-boxes identified to date in all species. A search of the *Drosophila* genome failed to identify any homologs to the mammalian p160 gene family of NR coactivator proteins. SMRTER, despite weak similarity to the mammalian corepressors SMRT and N-CoR, appears to be the only close relative in *Drosophila*.

Translation and RNA processing. Although the structure of the ribosome has been well worked out, it has become apparent that many ribosomal proteins are multifunctional and are involved in processes as disparate as DNA repair and iron-binding (*49*). There has been an enormous genetic investigation of the consequences of changes in expression level of *Drosophila* ribosomal proteins (the *Minute* phenotype) (*50*); the identification and mapping of the complete set presented here will provide the basis for in-depth dissections of their functions and disease roles.

Most genes encoding general translation factors are present in only one copy in the *Drosophila* genome, as they are in other genomes studied to date; however, we discovered six genes encoding proteins highly similar to the messenger RNA (mRNA) cap-binding protein eIF4E. These may add complexity to regulation of cap-dependent translation, which is central to cellular growth control. *Caenorhabditis elegans* has three eIF4E isoforms, which were hypothesized to be necessary because trans-spliced mRNAs possess a different cap structure than do other mRNAs (*51*); however, *Drosophila* does not have trans-spliced mRNAs. The activity of eIF4E is regulated by an inhibitor protein, 4E-BP. The *Drosophila* genome contains only a single gene encoding 4E-BP; in contrast, mammals have at least three 4E-BP isoforms but perhaps fewer eIF4E isoforms than do flies. Of the more than 200 RNA-binding proteins identified, the most frequent structural classes are RRM proteins (114), DEAD- or DExH-box helicases (58), and KH-domain proteins (31). This distribution is similar to that observed in the *C. elegans* genome. These structural motifs are sometimes found in proteins for which experimental evidence indicates a function in DNA, rather than RNA, binding. Overall, the translational machinery appears well conserved throughout the eukaryotes.

The process of nonsense-mediated decay (*52*), the accelerated decay of mRNAs that cannot be translated throughout their entire length, has been genetically characterized in yeast and *C. elegans* but not in *Drosophila*. We found homologs of UPF1/SMG-2, SMG-1, and SMG-7 in the *Drosophila* genome, indicating that this process is conserved in flies.

Of particular interest are genes for components of the minor, or U12, spliceosome (*53*). Such introns are known in mammals, *Drosophila*, and *Arabidopsis*, but not *C. elegans*. Using conservative criteria (including a perfect match to the U12 consensus 5′ splice site for nucleotides 2 to 7, TATCCT), we found one intron that appears to be of the U12 type per 1000 genes. As expected, the minor spliceosome snRNAs U12, U4atac, and U6atac are present in the *Drosophila* genome. However, neither U11 nor the U11-associated 35-kD protein (*54*) could be identified in the sequence. It is possible that these components of the minor spliceosome are less well conserved, or that the minor spliceosome in *Drosophila* does not contain them.

Cytochrome P450. The cytochrome P450 monooxygenases (CYPs) are a large and ancient superfamily of proteins that carry out multiple reactions to enable organisms to rid themselves of foreign compounds. Human CYP2D6, for example, influences the metabolism of beta blockers, antidepressants, antipsychotics, and codeine, and insect CYPs function in the synthesis or degradation of hormones and pheromones and in the metabolism of natural and synthetic toxins, including insecticides (*55*). We found 90 P450 fly genes, of which four are pseudogenes, a figure that is comparable to the 80 CYPs of *C. elegans*. These 90 genes, some of which are clustered, are divided among 25 families, five of which are found in Lepidoptera, Coleoptera, Hymenoptera, Orthoptera, and Isoptera. However, more than half of the 90 genes belong to only two families, CYP4 and CYP6, the former family shared with vertebrates. CYP51, used in making cholesterol in animals and related molecules in plants and fungi, is absent from both the fly and worm genomes; it is well known that the fly must obtain cholesterol from its diet. A comprehensive collection of phylogenetically diverse CYP sequences is available (*56*).

Solute transport. Solute transporters contribute to the most basic properties of living systems, such as establishment of cell potential or generation of ATP; in higher eukaryotes, these proteins help mediate advanced functions such as behavior, learning, and memory. Hydropathy analyses predict that 20% of the gene products in *Drosophila* reside in cellular membranes, having four or more hydrophobic α helices (*57*). A considerable fraction of these proteins (657, or 4%) are dedicated to ion and metabolite movement. More than 80% of the annotated transporters are new to *Drosophila* and were identified by similarity to proteins characterized in other eukaryotes. The largest families are sugar permeases, mitochondrial carrier proteins, and the ATP-binding cassette (ABC) transporters, with 97, 38, and 48 genes, respectively; these families are also the most common in yeast and *C. elegans* (*29*). Also of note are three families of anion transporters that mediate flux of sulfate, inorganic phosphate, and iodide. Na^+-anion transporters, with 17 members, are particularly abundant relative to worm and yeast. Although individual members of these families have been investigated—for example, the mitochondrial carrier protein COLT required for gas-filling of the tracheal system (*58*) and the ABC transporters associated with eye pigment distribution (*59*)—the variety and number of transporters within each family are impressive. These data lay the foundation for understanding global transport processes critical to *Drosophila* physiology and development.

Metabolic processes. The biosynthetic networks of the fly are remarkably complete compared to those of many different prokaryotes and to yeast, in which key enzymes of various pathways may be missing (*60*). As in vertebrates, many fly enzymes are encoded by multiple genes. Two families are noteworthy because of their size. The triacylglycerol lipases are encoded by 31 genes and merit consideration in investigations of lipolysis and energy storage and redistribution. In addition, there are 32 genes encoding uridine diphosphate (UDP) glycosyltransferases, which participate in the production of sterol glycosides and in the biodegradation of hydrophobic compounds. Several UDP glycosyltransferase genes are highly expressed in the antennae and may have roles in olfaction. In vertebrates, these enzymes are critical to drug clearance and detoxification (*61*). A major challenge will be to determine whether the number of these proteins present in the genome is correlated with the importance and complexity of the regulatory events involved in any given enzymatic reaction.

Iron (Fe) is both essential for and toxic to for all living things, and metazoan animals use similar strategies for obtaining, transporting, storing, and excreting iron. Three findings from the analysis of the genome shed light on the underlying common mechanisms that have escaped attention in the past. First, a third ferritin gene has been found that probably encodes a subunit belonging to a cytosolic ferritin, the predominant type in vertebrates. This finding indicates that intracellular iron storage mechanisms in flies might be very similar to those in vertebrates. Subunits of the

predominant secreted ferritins in insects are encoded by two highly expressed autosomal genes (*62*). Second, the dipteran transferrins studied so far appear to play antibiotic rather than iron-transport roles; one such transferrin was previously characterized in *Drosophila* (*63*). We have now identified two additional transferrins. The conservation of iron-binding residues and COOH-terminal hydrophobic sequences in these new transferrins suggests that they are homologs of the human melanotransferrin p97. The latter is anchored to the cells and mediates iron uptake independently from the main vertebrate pathway that involves serum transferrin and its receptor (*64*). Third, proteins homologous to vertebrate transferrin receptors appear to be absent from the fly. Thus, the *Drosophila* homologs of the vertebrate melanotransferrin could mediate the main insect pathway for cellular uptake of iron and possibly of other metal and nonmetal small ligands. This appears to be an ancestral mechanism, and the exploration of these findings should be crucial in bringing together what has seemed to be divergent iron homeostasis strategies in vertebrates and insects.

This initial look at the genomic basis of the fly's fundamental biochemical pathways reveals that its biosynthetic networks are fairly consistent with those of worm and human. On the other hand, there are a number of new findings. The large diversity of transcription factors, including several hundred zinc-finger proteins and novel homeodomain-containing proteins and nuclear hormone receptors, is likely related to the substantial regulatory complexity of the fly. In addition, many of the genes involved in core processes are single-copy genes and thus provide starting points for detailed studies of phenotype, free of the complications of genetically redundant relatives.

Fig. 4. Coding content of the fly genome. Each predicted gene in the genome is depicted as a box color-coded by similarity to genes from mammals, *C. elegans*, and *S. cerevisiae*. A legend appears at the end of each chromosome arm describing the components of each panel. In order from the top, they are (A) scale in megabases, (B) polytene chromosome divisions, (C) GC content in a range from 25 to 65%, (D) transposable elements, and genes on the (E) plus and (F) minus strands. The width of each gene element represents the total genomic length of the transcription unit. The height of each gene element represents EST coverage: The shortest boxes have no EST matches, medium-size boxes have 1 to 12 EST matches, and the tallest boxes have 13 or more EST matches. The color code for sequence similarity appears on each side of the fold-out figure. The graphics for this figure were prepared using gff2ps (*68*). Each gene has been assigned a FlyBase identifier (FBgn) in addition to the Celera identifier (CT#). Access to supporting information on each gene is available through FlyBase at http://flybase.bio.indiana.edu. These data are also available through a graphical viewing tool at FlyBase (http://flybase.bio.indiana.edu) and Celera (www.celera.com), with additional supporting information.

Concluding Remarks

Genome assembly relied on the use of several types of data, including clone-based sequence, whole-genome sequence from libraries with three insert sizes, and a BAC-based STS content map. The combination of these resources resulted in a set of ordered contigs spanning nearly all of the euchromatic region on each chromosome arm. We are taking advantage of the cloned DNA available from both the clone-based and whole-genome subclones to fill the gaps between contigs; 331 have been filled, and the remainder are in progress.

It is useful to consider the relative contributions of the various data types to the finished product with respect to how similar programs might be carried out in the future. The BAC end-sequences and STS content map provided the most informative long-range sequence-based information at the lowest cost. Both BAC ends and STS map were necessary to link scaffolds to chromosomal locations. A higher density of BAC end-sequences, from libraries produced with a larger diversity of restriction enzymes (or even from a random-shear library), would have resulted in larger scaffolds at lower shotgun sequence coverage; this is our primary recommendation for future projects. Although the clone-based draft sequence data did not result in a markedly different extent of scaffold coverage compared to assembly without the clone-based data, they were useful in the resolution of repeated sequences, particularly in the transition zones between euchromatin and centric heterochromatin. In terms of sequence coverage, adequate scaffold size was obtained with whole-genome sequence coverage as low as 6.5× (*11*). The assembly algorithm did not take any specific advantage of the fact that each draft sequence read from a BAC clone came from a defined region of the genome. Adding this feature could mean that adequate genome assembly could be obtained at lower whole-genome sequence coverage. Contiguity and scaffold size continued to increase with increased coverage, and so a decision to proceed with additional sequencing versus more directed gap closure should be driven by available resources.

The assembled sequence has allowed a first look at the overall *Drosophila* genome structure. As previously suspected, there is no clear boundary between euchromatin and heterochromatin. Rather, over a region of 1 Mb, there is a gradual increase in the density of transposable elements and other repeats, to the point that the sequence is nearly all repetitive. However, there are clearly genes within heterochromatin, and we suspect that most of our 3.8 Mb of unmapped scaffolds represent such genes, both near the centromeres and on the Y chromosome (which is almost entirely heterochromatic). Access to these sequences was an unexpected benefit of the WGS approach.

The genome sequence and the set of 13,601 predicted genes presented here are considered Release 1. Both will evolve over time as additional sequence gaps are closed, annotations are improved, cDNAs are sequenced, and genes are functionally characterized. The diversity of predicted genes and gene products will serve as the raw material for continued experimental work aimed at unraveling the molecular mechanisms underlying development, behavior, aging, and many other processes common to metazoans for which *Drosophila* is such an excellent model.

References and Notes

1. G. L. G. Miklos and G. M. Rubin, *Cell* **86**, 521 (1996).
2. A. S. Spradling *et al*., *Genetics* **153**, 135 (1999).
3. M. Ashburner *et al*., *Genetics* **153**, 179 (1999).
4. J. C. Venter *et al*., *Science* **280**, 1540 (1998).
5. G. M. Rubin and E. Lewis, *Science* **287**, 2216 (2000).
6. D. L. Hartl *et al*., *Trends Genet.* **8**, 70 (1992).
7. R. D. Fleischmann *et al*., *Science* **269**, 496 (1995); C. M. Fraser and R. D. Fleischmann, *Electrophoresis* **18**, 1207 (1997).
8. J. L. Weber and E. W. Myers, *Genome Res.* **7**, 409 (1997).
9. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
10. R. Hoskins *et al*., *Science* **287**, 2271 (2000).
11. E. W. Myers *et al*., *Science* **287**, 2196 (2000).
12. A number of methods were used to close gaps. Whenever possible, gaps were localized to a chromosome region and a spanning genomic clone was identified. When a spanning clone could be identified, it was used as a template for sequencing. The sequencing approach was determined by the gap size. For gaps smaller than 1 kb, BAC templates were sequenced directly with custom primers. For gaps larger than 1 kb, 3-kb plasmids or M13 clones from the clone-based draft sequencing were sequenced by directed methods, or 10-kb plasmids from the WGS sequencing project were sequenced by random transposon-based methods. If no 3-kb or 10-kb plasmid could be identified, PCR products were amplified from BAC clones or genomic DNA and end-sequenced directly with the PCR primers.
13. K. S. Weiler and B. T. Wakimoto, *Annu. Rev. Genet.* **29**, 577 (1995); S. Henikoff, *Biochem. Biophys. Acta* **1470**, 1 (2000); S. Pimpinelli *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3804 (1995); A. R. Lohe, A. J. Hilliker, P. A. Roberts, *Genetics* **134**, 1149 (1993).
14. G. L. G. Miklos, M. Yamamoto, J. Davies, V. Pirrotta, *Proc. Natl. Acad. Sci U.S.A.* **85**, 2051 (1988).
15. See ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/nuclear_cds_set.embl.v2.9.Z.
16. The genes found in unscaffolded sequence were Su(Ste) (FlyBase identifier FBgn0003582) on the Y chromosome, His1 (FBgn0001195) and His4 (FBgn0001200) (histone genes were screened out before assembly), rbp13 (FBgn0014016), and idr (FBgn0020850).
17. C. Burge and S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
18. M. G. Reese, D. Kulp, H. Tammana, D. Haussler, *Genome Res.*, in press.

19. Sequence contigs were searched against publicly available sequence at the DNA level and as six-frame translations against public protein sequence data. DNA searches were against the invertebrate (INV) division of GenBank, a set of 80,000 EST sequences produced at BDGP assembled to produce consensus sequences (*21*), and a set of curated *Drosophila* protein-coding genes prepared by three of the authors (M. Ashburner, L. Bayraktaroglu, and P. V. Benos) (*15*). Protein searches were performed against this set of curated protein sequences and against the nonredundant protein database available at the National Center for Biotechnology Information. Initial searches were performed with a version of BLAST2 (*25*), optimized for the Compaq Alpha architecture. Additional processing of each query-subject pair was performed to improve the alignments. All BLAST results having an expectation score of $<1 \times 10^{-4}$ were then processed on the basis of their high-scoring pair (HSP) coordinates on the contig to remove redundant hits, retaining hits that supported possible alternative splicing. This procedure was performed separately by hits to particular organisms so as not to exclude HSPs that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the contig sequence were then realigned to the contig with Sim4 [L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, W. Miller, *Genome Res*. **8**, 967 (1998)] for ESTs, and with Lap [X. Huang, M. D. Adams, H. Zhou, A. R. Kerlavage, *Genomics* **46**, 37 (1995)] for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually respect intron-exon boundaries and thus facilitate human annotation. Some regions of the genome may be underannotated because the bulk of the annotation work was done on an earlier assembly version. Continued updates will be available through FlyBase.
20. M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, S. E. Lewis, *Genome Res.*, in press.
21. G. M. Rubin *et al*., *Science* **287**, 2222 (2000).
22. See the Gene Ontology Web site (www.geneontology.org).
23. See the *Saccharomyces* Genome Database Web site (http://genome-www.stanford.edu/Saccharomyces).
24. D. Allen and J. Blake, Mouse Genome Informatics (www.informatics.jax.org).
25. S. F. Altschul *et al.*, *Nucleic Acids Res*. **25**, 3389 (1997).
26. S. M. Mount *et al*., *Nucleic Acids Res*. **20**, 4255 (1992).
27. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
28. X. Lin *et al*., *Nature* **402**, 761 (1999).
29. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
30. A. Dutta and S. P. Bell, *Annu. Rev. Cell Dev. Biol*. **13**, 293 (1997).
31. I. Chesnokov, M. Gossen, D. Remus, M. Botchan, *Genes Dev*. **13**, 1288 (1999).
32. G. Feger, *Gene* **227**, 149 (1999).
33. D. T. Pak *et al*., *Cell* **97**, 311 (1997); J. Rohrbough, S. Pinto, R. M. Mihalek, T. Tully, K. Broadie, *Neuron* **23**, 55 (1999).
34. S. Waga, G. J. Hannon, D. Beach, B. Stillman, *Nature* **369**, 574 (1994); H. Flores-Rozas *et al*., *Proc. Natl. Acad. Sci. U.S.A*. **91**, 8655 (1994).
35. R. Jessberger, C. Frei, S. M. Gasser, *Curr. Opin. Genet. Dev*. **8**, 254 (1998); T. Hirano, *Curr. Opin. Genet. Dev*. **10**, 317 (1998); A. V. Strunnikov, *Trends Cell Biol*. **8**, 454 (1998).
36. R. Saffery *et al*., *Hum. Mol. Genet.* **9**, 175 (2000); J. M. Craig, W. C. Earnshaw, P. Vagnarelli, *Exp. Cell Res*. **246**, 249 (1999); R. Saffery *et al*., *Chromosome Res*. **7**, 261 (1996).
37. R. Belotserkovskaya and S. L. Berger, *Crit. Rev. Eukaryotic Gene Expr*. **9**, 221 (1999).
38. J. A. Eisen, K. S. Sweder, P. C. Hanawalt, *Nucleic Acids Res*. **23**, 2715 (1995); K. J. Pollard and C. L. Peterson, *Bioessays* **20**, 771 (1998).
39. E. V. Koonin, S. Zhou, J. C. Lucchesi, *Nucleic Acids Res*. **23**, 4229 (1995).
40. F. Jeanmougin *et al*., *Trends Biochem. Sci.* **22**, 151 (1997); F. Winston and C. D. Allis, *Nature Struct. Biol*. **6**, 601 (1999).
41. R. W. Levis, *Mol. Gen. Genet.* **236**, 440 (1993); H. Biessmann and J. M. Mason, *Chromosoma* **106**, 63 (1997).
42. P. Gallinari and J. Jiricny, *Nature* **383**, 735 (1996).
43. B. Flores and W. Engels, *Proc. Natl. Acad. Sci. U.S.A*. **96**, 2964 (1999).
44. K. Kusano, M. E. Berres, W. R. Engels, *Genetics* **151**, 1027 (1999); J. J. Sekelsky, M. H. Brodsky, G. M. Rubin, R. S. Hawley, *Nucleic Acids Res*. **27**, 3762 (1999).
45. M. Hampsey, *Microbiol. Mol. Biol. Rev.* **62**, 465 (1998); R. H. Reeder, *Prog. Nucleic Acid Res. Mol. Biol*. **62**, 293 (1999); I. M. Willis, *Eur. J. Biochem*. **212**, 1 (1993).
46. T. I. Lee and R. A. Young, *Genes Dev*. **12**, 1398 (1998); M. Hampsey and D. Reinberg, *Curr. Opin. Genet. Dev*. **9**, 132 (1999).
47. M. D. Rabenstein, S. Zhou, J. T. Lis, R. Tjian, *Proc. Natl. Acad. Sci. U.S.A*. **96**, 4791 (1999).
48. D. Duboule, Ed., *Guidebook to the Homeobox Genes* (Oxford Univ. Press, New York, 1994).
49. I. G. Wool, *Trends Biochem. Sci*. **21**, 164 (1996).
50. A. Lambertsson, *Adv. Genet.* **38**, 69 (1998).
51. M. Jankowska-Anyszka *et al.*, *J. Biol. Chem*. **273**, 10538 (1998).
52. M. R. Culbertson, *Trends Genet*. **15**, 74 (1999).
53. C. Burge, T. Tuschl, P. Sharp, in *The RNA World*, R. Gesteland, T. Cech, J. Atkins, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1999).
54. C. L. Will, C. Schneider, R. Reed, R. Luhrmann, *Science* **284**, 2003 (1999).
55. R. Feyereisen, *Annu. Rev. Entomol*. **44**, 507 (1999).
56. See D. Nelson's Web site (http://drnelson.utmem.edu/CytochromeP450.html).
57. G. von Heijne, *J. Mol. Biol*. **225**, 487 (1992).
58. K. Hartenstein *et al*., *Genetics* **147**, 1755 (1997).
59. R. G. Tearle, J. M. Belote, M. McKeown, B. S. Baker, A. J. Howells, *Genetics* **122**, 595 (1989).
60. R. Maleszka, *Microbiology* **143**, 1781 (1997).
61. Q. Wang, G. Hasan, C. W. Pikielny, *J. Biol. Chem*. **274**, 10309 (1999).
62. B. C. Dunkov and T. Georgieva, *DNA Cell Biol*. **18**, 937 (1999).
63. T. Yoshiga *et al*., *Eur. J. Biochem.* **260**, 414 (1999).
64. M. L. Kennard *et al*., *EMBO J.* **14**, 4178 (1995).
65. High molecular weight genomic DNA was prepared from nuclei isolated [C. D. Shaffer, J. M. Wuller, S. C. R. Elgin, *Methods Cell Biol*. **44**, 185 (1994)] from 2.59 g of embryos of an isogenic *y; cn bw sp* strain [B. J. Brizuela *et al*., *Genetics* **137**, 803 (1994)]. The genomic DNA was randomly sheared, end-polished with Bal31 nuclease/T4 DNA polymerase, and carefully size-selected on 1% low-melting-point agarose. After ligation to BstX1 adaptors, genomic fragments were inserted into BstX1-linearized plasmid vector. Libraries of 1.8 ± 0.2 kb were cloned in a high-copy pUC18 derivative, and libraries of 9.8 ± 1.0, 10.5 ± 1.0, and 11.5 ± 1.0 kbp were cloned in a medium-copy pBR322 derivative. High-throughput methods in 384-well format were implemented for plasmid growth, alkaline lysis plasmid purification, and ABI Big Dye Terminator DNA sequencing reactions. Sequence reads from the genomic libraries were generated over a 4-month period using 300 DNA analyzers (ABI Prism 3700). These reads represent more than 12× coverage of the 120-Mbp euchromatic portion of the *Drosophila* genome (Table 1). Base-calling was performed using 3700 Data Collection (PE Biosystems) and sequence data were transferred to a Unix computer environment for further processing. Error probabilities were assigned to each base with TraceTuner software developed at Paracel Inc. (www.paracel.com). The predicted error probability was used to trim each sequence read such that the overall accuracy of each trimmed read was predicted to be >98.5% and no single 50-bp region was less than 97% accurate. The efficacy of TraceTuner and the trimming algorithm was demonstrated by comparing trimmed sequence reads to high-quality finished sequence data from BDGP (Fig. 2).
66. For clone-based genomic sequencing, BAC, P1, and cosmid DNAs were prepared by alkaline lysis procedures and purified by CsCl gradient ultracentrifugation. DNA was randomly sheared and size-selected on LMP agarose for fragments in the 3-kb range for plasmids and in the 2-kb range for M13 clones. After blunt-ending with T4 DNA polymerase, plasmids were generated by ligation to BstX1 adaptors and insertion into BstX1-linearized pOT2A vector. M13 clones were generated using the double-adaptor protocol [B. Andersson *et al*., *Anal. Biochem*. **236**, 107 (1996)]. Plasmid sequencing templates were prepared by alkaline lysis (Qiagen) or by PCR, and M13 templates were prepared using the sodium perchlorate–glass fiber filter technique [B. Andersson *et al*., *Biotechniques* **20**, 1022 (1996)]. Paired end-sequences of 3-kb plasmid subclones were generated (principally) with ABI Big Dye Terminator chemistry on ABI 377 slab gel or ABI 3700 capillary sequencers. Additional M13 subclone sequence was generated using BODIPY dye-labeled primers. Procedures for finishing sequence to high quality at LBNL were as described (*3*).
67. M.-T. Yamamoto *et al*., *Genetics* **125**, 821 (1990).
68. J. F. Abril and R. Guigo, *Bioinformatics*, in press.
69. A. Peter *et al*., in preparation.
70. J. Locke, L. Podemski, N. Aippersbach, H. Kemp, R. Hodgetts, in preparation.

71. The many participants from academic institutions are grateful for their various sources of support. We thank B. Thompson and his staff for the excellent laboratories and work environment, M. Peterson and his team for computational support, and V. Di Francesco, S. Levy, K. Chaturvedi, D. Rusch, C. Yan, and V. Bonazzi for technical discussions and thoughtful advice. We are indebted to R. Guigo and to E. Lerner of Aquent Partners for assistance with illustrations. The work described was funded by Celera Genomics, the Howard Hughes Medical Institute, and NIH grant P50-HG00750 (G.M.R.).